Measures for the exceptionality of gene order in conserved genomic regions

Simona Grusea

Measures for the exceptionality of gene order in conserved genomic regions

Simona Grusea[1]

LATP – UMR CNRS 6632, Équipe EBM, Université de Provence,
Institut de Mathématiques de Toulouse, INSA de Toulouse, Université de Toulouse

Abstract

We propose in this article three measures for quantifying the exceptionality of gene order in conserved genomic regions found by the reference region approach. The three measures are based on the transposition distance in the permutation group. We obtain analytic expressions for their distribution in the case of a random uniform permutation, i.e. under the null hypothesis of random gene order. Our results can be used to increase the power of the significance tests for gene clusters which take into account only the proximity of the orthologous genes and not their order.

AMS 2000 subject classifications: Primary: 60C05; secondary: 92D15.

Key words: gene order comparison, distance between permutations, transposition distance, exact distribution, random uniform permutation, conserved genomic regions, reference region approach.

1. Introduction

1.1. The biological context

Comparing genomic organization between different species may help to decipher the evolutionary history of species and also to better understand the biology of nowadays species.

Orthologous genes are two genes, in two different species, that descend from the same gene at the ancestor of the two species, as the result of a speciation event. They tend, in general, to have similar functions. Therefore, finding a group of orthologous genes in close proximity in the genomes of two different species may be the mark of evolutionary relatedness between the two species or the sign of a functional relationship among these genes, together with a functional selective pressure acting on them. But for this to be the case, the observed

Email address: gruseas@yahoo.com (Simona Grusea)
[1]Address: INSA de Toulouse, Département GMM, 135 avenue de Rangueil - 31077 Toulouse, France.

orthologous gene clusters between the two genomes have to be *significant*, i.e. very improbable to have appeared by chance.

Over evolutionary time scales, the gene order in one genome can be affected by various genome rearrangement events, like inversions, translocations, transpositions, chromosomic fissions and fusions. Hence, in the absence of certain constraints due to functional selective pressures, the gene order is rapidly randomized. This is one reason why, in general, the null hypothesis taken in the significance tests for gene clusters is the hypothesis of random gene order.

In the literature various definitions for gene clusters exist, and also different statistical tests for detecting gene clusters which are significant from the point of view of the proximity of the orthologs (see [2, 6, 9, 12, 13, 17]).

Here we choose a very unrestrictive definition for a gene cluster. We call *conserved genomic region* or *gene cluster* two chromosomic regions, in two different species, that have in common a certain number of orthologous genes, not necessarily adjacent or in the same order in the two genomes. We do not impose any restriction on the gap length between consecutive orthologs.

Suppose that we have found an orthologous gene cluster which is significant from the point of view of the proximity of the genes. One might want to take into account also the order of the orthologs in this cluster, considering that the gene clusters in which the gene order is more similar are even more biologically significant.

The goal of this work is to find "good" measures for quantifying the degree of conservation of the order of the orthologs in conserved genomic regions. Here, "good" means both biologically relevant and computationally accessible.

We present in this article three measures based on the transposition distance in the permutation group, together with analytic expressions for their distributions under the null hypothesis of random gene order. Our results may serve as a tool to increase the power of the statistical tests for detecting significant gene clusters.

There are different approaches when searching for gene clusters (see [9]). In this article we focus on the case where the gene clusters were found by the "reference region" approach, which consists in starting with a fixed genomic region in a certain species A, called *the reference region*, and searching for significant orthologous gene clusters in the whole genome of another species B.

A genome will be seen as an ordered set of genes. In this article we are only considering the case of genomes with a single chromosome. Also, we are treating only the case of no multigene families, i.e. we suppose that each gene from the reference region in A has at most one ortholog in the genome B.

1.2. Previous work

The literature on this subject is just at its beginnings. Sankoff and Haque [18] propose three adjacency disruption measures for comparing the order of the orthologs which are in common between two clusters in two genomes. They investigate in more detail the "maximum adjacency disruption" criterion, giving analytic formulas for some values of its distribution under random gene order

and also simulation results. They also note the difficulty of taking into account, in a single statistical test, both the proximity of the orthologs and their order.

Recently, Xu and Sankoff [20] used a parametrized definition of gene clusters based on the notion of generalized adjacency, which allows to take into account gene content and gene order at the same time. For cluster sizes smaller than 4, they derived exact values for the expected number of clusters. For larger cluster sizes, they used simulations to study the distribution of the number of clusters.

In the "genome rearrangements" literature, several genomic distances have been studied, which take into account one or a combination of different types of genomic events: reversals, translocations, chromosomic fissions and fusions, biological transpositions, block-interchanges – see [15] for a review.

The problem with using these distances as test statistics comes from the fact that their distribution for a random uniform permutation is in general unknown and there are very few theoretical results on this subject. Recently, Doignon and Labarre [8] have found the distribution of the number of alternating cycles in the bicolored breakpoint graph of a random (unsigned) permutation, which can be used to deduce the exact distribution of the "block-interchange" distance of Christie [4]. Sankoff and Haque [19] and Xu, Zheng and Sankoff [21], using a constructive approach, have obtained asymptotic estimates for the distribution of the number of cycles in the breakpoint graph of two random signed permutations, which can be further used to estimate the distribution of the genomic distances based on the breakpoint graph.

1.3. The transposition distance

All the three measures that we propose in this article are based on the mathematical transposition distance in the permutation group.

A mathematical transposition applied to a permutation exchanges two elements (not necessarily adjacent) in the permutation. A rigorous definition will be given in the next section.

The genomic equivalent operation would be the exchange of two genes (not necessarily adjacent) in the genome. Therefore, the mathematical transpositions do not model a real biological operation.

The transposition distance is very "nice" from the computational point of view, but could it also be useful in the biological context?

One positive answer to this question is given by Eriksen and Hultman in [10], where they describe an analogy between mathematical transpositions and genomic reversals. They show that the expected transposition distance in S_n after applying t random transpositions to the identity is a very good approximation for the expected reversal distance of a genome with n genes (seen as a signed permutation) after applying t random reversals to the identity. By obtaining a closed formula for the first, they propose an estimate by the method of moments for the true evolutionary distance between two genomes. They show that their method compares well to the best results obtained through other methods.

Similarities between the distributions of the two distances are also discussed in [1].

The paper is organized as follows. In Section 2 we present the mathematical framework and formulate the problem. In Sections 3, 4 and 5 we propose three measures for the degree of conservation of gene order in a conserved genomic region, and for each of these measures we derive its exact distribution in the case of a random uniform permutation. Section 6 is dedicated to some concluding remarks and ideas for future work.

We think that the originality of the measures presented in Sections 4 and 5 comes from the fact that they are taking into account not only the order of the genes which are in common between the two clusters, but also the positions of the other orthologs. These measures are specifically adapted to the case where the gene clusters are found by the reference region approach.

We hope that the results obtained in this article are interesting not only in the genomic comparison context, but also in themselves, as a modest contribution to the giant pool of results about the transposition distance in the permutation group.

2. The mathematical framework

Let n denote the number of genes in the reference region in A which have one and only one ortholog in B.

The null hypothesis that we consider is

$$H_0 : \text{random gene order in the genome B,}$$

under which all the $n!$ possible orderings of the n orthologs have the same probability to occur.

We label the n orthologs in such a way that their order in the reference region is given by the identity permutation Id_n, and we let π denote the permutation representing their order in the genome B.

Under the null hypothesis of random gene order in the genome B, π is a random permutation of n elements, uniformly chosen with probability $1/n!$.

For a permutation $\pi \in S_n$, we will use the notation $\pi = [\pi(1), ..., \pi(n)]$.

Suppose that we are interested in the gene order in a certain conserved genomic region $\mathcal{R}$ in the genome B, which contains h orthologs and starts with the i-th ortholog, labelled $\pi(i)$.

In what follows h and i will be fixed.

We would like to find a way to quantify the conservation of the gene order in the region $\mathcal{R}$ compared to the order of the genes in the reference region.

Notation 1. For a permutation $\sigma \in S_n$ and for $i \in \{1, ..., n-h+1\}$, we denote by $\sigma_{i,h}$ the restriction of σ to the set $\{i, i+1, ..., i+h-1\}$, i.e.

$$\sigma_{i,h} = [\sigma(i), ..., \sigma(i+h-1)].$$

Note that $\pi_{i,h}$ contains the labels of the h orthologs from the region of interest $\mathcal{R}$.

For comparing two permutations we will use the transposition distance in the symmetric group S_n, which we denote d_{trp}.

We recall that a *transposition* in the group S_n is a cycle of length 2. The composition to the right of a given permutation $\pi = [\pi(1), ..., \pi(n)]$ with a transposition (i, j) results in the permutation

$$\pi \circ (i, j) = [\pi(1), ..., \pi(i-1), \pi(j), \pi(i+1), ..., \pi(j-1), \pi(i), \pi(j+1), ..., \pi(n)],$$

in which the elements $\pi(i)$ and $\pi(j)$ are interchanged.

The permutation group S_n is generated by the set of all the transpositions.

For two permutations $\pi, \sigma \in S_n$, the transposition distance $d_{trp}(\pi, \sigma)$ is the minimum number of transpositions needed to transform π into σ or conversely.

We will use the following two classical results about permutations (see [7], p. 118 for the first one and [5], p. 234 for the second one).

Lemma 1. *If σ is a permutation of n elements, then*

$$d_{trp}(\sigma, Id_n) = n - c(\sigma),$$

where $c(\sigma)$ denotes the number of cycles in the disjoint cycle decomposition of σ, counting also the singletons.

Lemma 2. *The number of permutations of n elements which have k disjoint cycles is given by the* signless Stirling number of the first kind *$s(n, k)$ (see Definition 1(ii)).*

Definition 1. We have the following equivalent definitions for the *signless Stirling number of the first kind*, which we denote $s(n, k)$ (see [5], chapter V):

(*i*) $s(n, k)$, $n \geq k \geq 0$ satisfy the recurrence relation

$$s(n+1, k) = ns(n, k) + s(n, k-1), \tag{1}$$

with boundary conditions $s(0, 0) = 1$ and $s(n, 0) = 0$ for $n > 0$.

(*ii*) For $n \geq k \geq 1$, $s(n, k)$ equals the number of permutations of n elements which have k cycles in the disjoint cycle decomposition.

(*iii*) For $n \geq k \geq 1$, $s(n, k)$ is the coefficient of the term x^k in the expansion of $x_{(n)} = x(x+1)\cdots(x+n-1)$.

3. A first distance

A first idea, and the most simple one, is to compare only the order of the h orthologous genes in the region $\mathcal{R}$, given by $\pi_{i,h}$, with their order in the reference region, given by the restriction $Id_n|_{\{\pi(i),...,\pi(i+h-1)\}}$.

Note that under the null hypothesis, $\pi_{i,h}$ is a random permutation of h elements chosen among n and hence, for $0 \leq d \leq h-1$, we have

$$\mathbb{P}(d_{trp}(\pi_{i,h}, Id_n|_{\{\pi(i),...,\pi(i+h-1)\}}) \leq d) = \mathbb{P}(d_{trp}(\sigma, Id_h) \leq d),$$

where σ is a random permutation of h elements.

Notation 2. For $1 \leq k \leq n$, we will denote by

$$p(n,k) := \frac{1}{n!} \sum_{j=k}^{n} s(n,j)$$

the probability that a random permutation of n elements has at least k cycles in its disjoint cycle decomposition.

Using Lemma 1 and Lemma 2, we obtain

Proposition 3. *For $0 \leq d \leq h-1$, we have*

$$\mathbb{P}(d_{trp}(\pi_{i,h}, Id_n|_{\{\pi(i),\ldots,\pi(i+h-1)\}}) \leq d) = p(h, h-d).$$

Notice that the distance considered here takes into account only the relative order of the h orthologs which are common to the reference region and to the orthologous region $\mathcal{R}$, and it ignores the positions of the other orthologs.

4. A second distance

Because we are interested in measuring the conservation of the gene order between the reference region and the region $\mathcal{R}$, another idea is to still ignore the order of the other $n-h$ orthologs in the genome B, but to take into account their positioning in the reference region.

We consider the following "distance":

$$d(\pi, Id_n) := \min\{d_{trp}(\sigma, Id_n) : \sigma \in S_n, \sigma_{i,h} = \pi_{i,h}\},$$

i.e. we take the minimum over all possible orderings in B of the other orthologs outside the region $\mathcal{R}$.

Note that $d(\pi, Id_n)$ depends only on $\pi_{i,h}$ and not on the whole permutation π.

Consequently, d is not a real distance from the mathematical point of view, as we can have $d(\pi, Id_n) = 0$ without necessarily having $\pi = Id_n$. For $d(\pi, Id_n)$ to vanish it is sufficient to have $\pi_{i,h} = [i, ..., i+h-1]$.

In order to stress the fact that $d(\pi, Id_n)$ depends only on $\pi_{i,h}$, in what follows we will denote it $d(\pi_{i,h}, Id_n)$.

We have the following result.

Proposition 4.

$$d(\pi_{i,h}, Id_n) = h - cc(\pi_{i,h}),$$

*where $cc(\pi_{i,h})$ denotes the number of cycles of π which contain only elements belonging to $\{i, ..., i+h-1\}$. We will call these cycles "*closed cycles of $\pi_{i,h}$*".*

Proof. Indeed, by Lemma 1, we have

$$d(\pi_{i,h}, Id_n) = n - \max\{c(\sigma) : \sigma \in S_n, \sigma_{i,h} = \pi_{i,h}\},$$

and it suffices to note that the maximum is attained by the unique permutation σ° verifying $\sigma^\circ_{i,h} = \pi_{i,h}$ and for which the elements from $\{1, ..., n\} \backslash \{i, ..., i+h-1\}$ are all in distinct cycles. The permutation σ° has exactly $cc(\pi_{i,h}) + n - h$ cycles and hence

$$d(\pi_{i,h}, Id_n) = n - c(\sigma^\circ) = h - cc(\pi_{i,h}). \qquad \square$$

In the next theorem we give the distribution of $d(\pi_{i,h}, Id_n)$ under the null hypothesis.

Theorem 5. *For $0 \leq d \leq h-1$, we have*

$$\mathbb{P}(d(\pi_{i,h}, Id_n) \leq d) = \frac{1}{\binom{n}{h}} \sum_{m=h-d}^{h} \binom{n-m-1}{n-h-1} p(m, h-d).$$

Proof. Let M denote the r.v. representing the number of elements from $\{i, ..., i+h-1\}$ which are included in closed cycles of $\pi_{i,h}$.

We will compute $\mathbb{P}(d(\pi_{i,h}, Id_n) \leq d)$ by conditioning on the values of M:

$$\mathbb{P}(d(\pi_{i,h}, Id_n) \leq d) = \sum_{m=h-d}^{h} \mathbb{P}(d(\pi_{i,h}, Id_n) \leq d | M = m) \mathbb{P}(M = m). \tag{2}$$

The key observation is that the conditional distribution of $cc(\pi_{i,h})$ given $M = m$ is the same as the distribution of the number of cycles in a random permutation of m elements. Hence

$$\mathbb{P}(d(\pi_{i,h}, Id_n) \leq d | M = m) = p(m, h-d). \tag{3}$$

It remains to find the distribution of M.

Recall that, under H_0, $\pi_{i,h}$ is a random permutation of h elements chosen among n. So, for $h - d \leq m \leq h$:

$$\mathbb{P}(M = m) = \frac{|\{\pi_{i,h} : M = m\}|}{h! \binom{n}{h}}. \tag{4}$$

Every $\pi_{i,h}$ determines a unique permutation σ° such that $\sigma^\circ_{i,h} = \pi_{i,h}$ and having all the elements from $\{1, ..., n\} \setminus \{i, ..., i+h-1\}$ in distinct cycles.

Hence $|\{\pi_{i,h} : M = m\}|$ equals the number of permutations in S_n having all the elements from $\{1, ..., n\} \setminus \{i, ..., i+h-1\}$ in $n-h$ distinct cycles and such that exactly m elements among $i, i+1, ..., i+h-1$ do not belong to any of those cycles.

We then obtain

$$\begin{aligned} |\{\pi_{i,h} : M = m\}| &= \sum_{\substack{k_1, ..., k_{n-h} \geq 0 \\ k_1 + \cdots + k_{n-h} = h-m}} \binom{h}{k_1, ..., k_{n-h}, m} m! \prod_{j=1}^{n-h} k_j! \qquad (5) \\ &= h! \binom{n-m-1}{n-h-1}, \end{aligned}$$

where $k_1, ..., k_{n-h}$ count the number of elements from $\{i, i+1, ..., i+h-1\}$ which are in the cycles determined, respectively, by each of the elements in $\{1, ..., n\} \setminus \{i, ..., i+h-1\}$. The multinomial coefficient stands for the number of choices for those $k_1 + \cdots + k_{n-h}$ elements. The product of factorials counts the number of different ways of forming the $n-h$ cycles. For example, the cycle number j contains k_j+1 elements and there are $k_j!$ different cyclic permutations of those elements. The $m!$ term represents the number of ways of permuting the remaining elements from $\{i, i+1, ..., i+h-1\}$, elements which will form the closed cycles of $\pi_{i,h}$.

The last equality follows from the identity:

$$\left| \left\{ (k_1, ..., k_\ell) : k_j \geq 0, \forall j, \sum_{j=1}^{\ell} k_j = s \right\} \right| = \binom{s+\ell-1}{\ell-1}. \tag{6}$$

To prove this identity, we first make the change of variables $k'_j := k_j + 1$, $j = 1, ..., \ell$ and then we use the "bars and stars" idea to notice that

$$\left| \left\{ (k'_1, ..., k'_\ell) : k'_j \geq 1, \forall j, \sum_{j=1}^{\ell} k'_j = s+\ell \right\} \right|$$

represents the number of ways of separating $s+\ell$ stars arranged on a line, into ℓ nonempty groups. This number is exactly the binomial coefficient from the right-hand side of (5), because one has to place $\ell-1$ bars in $\ell-1$ of $s+\ell-1$ places available.

From (3) and (4) we deduce that for every $m \in \{h-d, ..., h\}$:

$$\mathbb{P}(M = m) = \frac{\binom{n-m-1}{n-h-1}}{\binom{n}{h}}. \tag{7}$$

By substituting (6) and (2) into (1), the formula in the statement of the theorem follows. □

5. A third distance

From the biological point of view, a disadvantage of the previous distance is the fact that it is very restrictive with respect to the position of the cluster $\mathcal{R}$ in the genome B.

For taking into account eventual genomic transpositions or translocations that could have changed the position of $\mathcal{R}$ with respect to the other orthologs in B, we think that a better idea would be to use the following "distance":

$$d^*(\pi, Id_n) := \min\{d_{trp}(\sigma, Id_n) : \sigma \in S_n, \sigma_{i^*,h} = \pi_{i,h}\}, \tag{8}$$

where

$$i^* := \arg \max_{1 \leq j \leq n-h+1} \left| \{\pi_i, ..., \pi_{i+h-1}\} \cap \{j, ..., j+h-1\} \right|.$$

We need to make a convention for the case when we have more than one maximum point. We decide, for example, to assign i^* the smallest such value.

As in the case of d, $d^*(\pi, Id_n)$ depends only on $\pi_{i,h}$ and not on the whole permutation π, hence d^* is not a real distance from the mathematical point of view. For $d^*(\pi, Id_n)$ to vanish it suffices to have $\pi_{i,h} = [j, ..., j+h-1]$, for some $j \in \{1, ..., n-h+1\}$.

In what follows we will denote $d^*(\pi, Id_n)$ by $d^*(\pi_{i,h}, Id_n)$.

We denote by σ° the unique permutation which attains the minimum in (7), hence has all the elements from $\{1, ..., n\} \setminus \{i^*, ..., i^*+h-1\}$ in distinct cycles.

As in Proposition 4, we have

$$d^*(\pi_{i,h}, Id_n) = h - cc(\sigma^\circ_{i^*,h}),$$

where $cc(\sigma^\circ_{i^*,h})$ denotes the number of "closed cycles of $\sigma^\circ_{i^*,h}$", i.e. the number of those cycles of σ° which do not contain any elements from $\{1, ..., n\} \setminus \{i^*, ..., i^*+h-1\}$.

Let

$$L^* := |\{\pi_i, ..., \pi_{i+h-1}\} \cap \{i^*, ..., i^*+h-1\}|.$$

Let $0 \leq d \leq h-1$. For computing the probability $\mathbb{P}(d^*(\pi_{i,h}, Id_n) \leq d)$ we will condition on the values of L^*:

$$\mathbb{P}(d^*(\pi_{i,h}, Id_n) \leq d) = \sum_{\ell=h-d}^{h} \mathbb{P}(d^*(\pi_{i,h}, Id_n) \leq d | L^* = \ell)\mathbb{P}(L^* = \ell).$$

Note that L^* must be greater than $h-d$, because we need to have at least $h-d$ closed cycles of $\sigma^\circ_{i^*,h}$ and hence at least $h-d$ elements in common between $\{\sigma^\circ_{i^*}, ..., \sigma^\circ_{i^*+h-1}\}$ and $\{i^*, ..., i^*+h-1\}$.

Next we will compute the conditional probabilities. We will prove:

Proposition 6. *For $0 \leq d \leq h-1$ and $h-d \leq \ell \leq h$, we have*

$$\mathbb{P}(d^*(\pi_{i,h}, Id_n) \leq d | L^* = \ell) = \frac{1}{\binom{h}{\ell}} \sum_{m=h-d}^{\ell} \binom{h-m-1}{h-\ell-1} p(m, h-d).$$

Proof. We denote by M^* the number of elements from $\{i^*, ..., i^*+h-1\}$ which are included in closed cycles of $\sigma^\circ_{i^*,h}$.

By further conditioning on M^* we obtain:

$$\mathbb{P}(d^*(\pi_{i,h}, Id_n) \leq d | L^* = \ell) = \sum_{m=h-d}^{\ell} \mathbb{P}(M^* = m | L^* = \ell) p(m, h-d). \tag{9}$$

Indeed, we have

$$\begin{aligned}\mathbb{P}(d^*(\pi_{i,h}, Id_n) \leq d | L^* = \ell, M^* = m) &= \mathbb{P}(d^*(\pi_{i,h}, Id_n) \leq d | M^* = m) \\ &= p(m, h-d).\end{aligned}$$

Let $\mathcal{P} = \{\pi_i, ..., \pi_{i+h-1}\}$. Under the null hypothesis, $\mathcal{P}$ is a random combination of h elements out of n.

Notice that i^* and L^* are completely determined by $\mathcal{P}$ and that M^* is completely determined given a permutation of $\mathcal{P}$.

We have

$$\mathbb{P}(M^* = m|L^* = \ell) = \sum_{\mathcal{P}:L^*=\ell} \mathbb{P}(M^* = m|\mathcal{P})\mathbb{P}(\mathcal{P}|L^* = \ell). \tag{10}$$

Fix $\mathcal{P}$ s.t. $L^* = \ell$. Then $\mathbb{P}(M^* = m\ |\mathcal{P})$ equals the number of permutations of $\mathcal{P}$ s.t. $M^* = m$ divided by $h!$.

Let $\mathcal{I} := \{i^*, ..., i^* + h - 1\} \setminus \mathcal{P}$ and $\mathcal{J} := \mathcal{P} \setminus \{i^*, ..., i^* + h - 1\}$. Both $\mathcal{I}$ and $\mathcal{J}$ have $h - \ell$ elements.

Notice that the number of permutations of $\mathcal{P}$ for which $M^* = m$ equals the number of permutations σ° in S_n verifying the following conditions:

(i) every element from $\mathcal{I}$ is in a cycle with exactly one element from $\mathcal{J}$;

(ii) there are no two elements of $\mathcal{I}$ or two elements of $\mathcal{J}$ in the same cycle;

(iii) all the elements from $\{1, ..., n\} \setminus (\mathcal{P} \cup \mathcal{I} \cup \mathcal{J})$ are fixed points;

(iv) there are exactly m elements from $\mathcal{P} \cap \{i^*, ..., i^* + h - 1\}$ which do not belong to any of the $h - \ell$ cycles determined by the $h - \ell$ pairs formed by one element from $\mathcal{I}$ and one element from $\mathcal{J}$;

(v) $\sigma^\circ(\mathcal{J}) = \mathcal{I}$.

We obtain

$$\begin{aligned}\mathbb{P}(M^* = m|\mathcal{P}) &= \frac{1}{h!}(h-\ell)! \sum_{\substack{k_1,...,k_{h-\ell}\geq 0 \\ k_1+\cdots+k_{h-\ell}=\ell-m}} \binom{\ell}{k_1, ..., k_{h-\ell}, m} m! \prod_{j=1}^{h-\ell} k_j! \qquad (11)\\ &= \frac{\binom{h-m-1}{h-\ell-1}}{\binom{h}{\ell}} \text{ (using (5))}.\end{aligned}$$

Indeed, we have $(h - \ell)!$ ways of pairing the elements from $\mathcal{I}$ with the elements from $\mathcal{J}$, and each of these pairings determines $h-\ell$ disjoint cycles. In the above formula, $k_1, ..., k_{h-\ell}$ denote the number of elements from $\mathcal{P} \cap \{i^*, ..., i^* + h - 1\}$ which belong to those $h - \ell$ cycles, respectively. The multinomial coefficient stands for the choice of these elements and the product of factorials counts the number of possibilities for forming the cycles. Consider, for example, the first cycle and denote by a and b its elements from $\mathcal{I}$ and $\mathcal{J}$, respectively. This cycle contains $k_1 + 2$ elements, but we have the restriction $\sigma^\circ(b) = a$ and hence we have only $k_1!$ ways of forming it. The $m!$ term counts the number of ways of permuting the m elements which form the closed cycles of $\sigma^\circ_{i^*,h}$.

Note that the formula (10) is the same for all $\mathcal{P}$ satisfying $L^* = \ell$ and hence, from (9), we deduce

$$\mathbb{P}(M^* = m | L^* = \ell) = \frac{\binom{h-m-1}{h-\ell-1}}{\binom{h}{\ell}}.$$

Substituting into (8), the formula in the statement follows. □

It remains to find the distribution of

$$L^* = \max_{1 \leq j \leq n-h+1} |\mathcal{P} \cap \{j, ..., j+h-1\}|,$$

where $\mathcal{P} = \{\pi_i, ..., \pi_{i+h-1}\}$ is a random combination of h elements among n.

Note that we can associate to each $\mathcal{P}$ a unique sequence $\boldsymbol{x}$ of zeros and ones, of length n, in the following manner: for every $k = 1, ..., n$, we let $x_k = 1$ if and only if $k \in \mathcal{P}$. Hence, under the null hypothesis of random gene order, $\boldsymbol{x}$ is a random sequence formed by h ones and $n - h$ zeros.

Therefore, L^* counts the maximum number of 1's within any window of length h in a random sequence formed with h 1's and $(n-h)$ 0's. This variable appears in the literature on scan statistics and it is called *discrete conditional scan statistic*, as it is a scan statistic in the case of i.i.d. Bernoulli r.v.'s, conditional on the number of successes (1's).

Several exact formulas for the distribution of the conditional discrete scan statistic exist in the literature, for different particular cases for the parameters, and also various approximations and bounds (see [11], chapter 12).

Here we will derive an exact expression for its distribution in the most general case, by adapting the results obtained by Huntington and Naus [14] in the conditional continuous settings. We have not seen this result in the literature, although it might have already appeared. We now give a proof of it. We follow the ideas from the proof of Huntington and Naus [14] and use a result of Naus [16].

Let $X_j, j = 1, ..., n$ be i.i.d. Bernoulli random variables.

Notation 3. For $1 \leq k \leq n - j + 1, j = 1, ..., n$, we denote

$$X_{j,k} := X_j + \cdots + X_{j+k-1}.$$

Let

$$N_h := \max_{1 \leq i \leq n-h+1} X_{i,h}.$$

Notice that

$$\mathbb{P}(L^* = \ell) = \mathbb{P}(N_h = \ell | X_{1,n} = h). \tag{12}$$

Let $2 \leq a \leq n$. We will give the result in the general settings where we condition on having a successes (1's). We have the following result.

Proposition 7. *If we denote by L the integer part of $\frac{n}{h}$ and we let $b = n - Lh$, then, for $2 \leq k \leq a$:*

$$\mathbb{P}(N_h < k | X_{1,n} = a) = \frac{(b!)^{L+1}[(h-b)!]^L}{\binom{n}{a}} \sum_{Q_k} \det |d_{ij}^{(k)}| \det |g_{ij}^{(k)}|,$$

where

$$Q_k = \{(n_1, ..., n_{2L+1}) : n_i \in \mathbb{N}, \textstyle\sum_{i=1}^{2L+1} n_i = a,\ n_i + n_{i+1} < k, \forall i = 1, ..., 2L\}$$

and the determinants are of size $(L+1) \times (L+1)$ and $L \times L$ respectively, with

$$d_{ij}^{(k)} = \frac{1}{c_{ij}^{(k)}!(b - c_{ij}^{(k)})!},\ g_{ij}^{(k)} = \frac{1}{f_{ij}^{(k)}!(h - b - f_{ij}^{(k)})!},$$

where

$$c_{ij}^{(k)} = \begin{cases} -\sum_{s=2i}^{2j-2} n_s + (j-i)k, & \text{if } 1 \leq i \leq j \leq L+1 \\ \sum_{s=2j-1}^{2i-1} n_s - (i-j)k, & \text{if } 1 \leq j < i \leq L+1 \end{cases}$$

and

$$f_{ij}^{(k)} = \begin{cases} -\sum_{s=2i+1}^{2j-1} n_s + (j-i)k, & \text{if } 1 \leq i \leq j \leq L \\ \sum_{s=2j}^{2i} n_s - (i-j)k, & \text{if } 1 \leq j < i \leq L. \end{cases}$$

Proof. We divide the n Bernoulli trials into $2L+1$ groups, the odd-numbered groups, $I_{2i-1}, i = 1, ..., L+1$, being of size b and the other ones, $I_{2i}, i = 1, ..., L$, of size $h - b$:

$$I_{2i-1} = \{(i-1)h + 1, ..., (i-1)h + b\}, i = 1, ..., L+1$$
$$I_{2i} = \{(i-1)h + b + 1, ..., ih\},\ i = 1, ..., L.$$

For $i = 1, ..., 2L+1$, we will denote by n_i the number of 1's in the i-th group, i.e.

$$n_i = \sum_{j \in I_i} X_j.$$

Conditional on $X_{1,n} = a$, the joint distribution of the n_i's is:

$$\mathbb{P}(n_1, ..., n_{2L+1} | X_{1,n} = a) = \frac{\prod_{i=1}^{L+1} \binom{b}{n_{2i-1}} \prod_{i=1}^{L} \binom{h-b}{n_{2i}}}{\binom{n}{a}},\ \text{if } \sum_{i=1}^{2L+1} n_i = a. \tag{13}$$

We denote

$$S_1 = \bigcup_{i=1}^{L+1} I_{2i-1},\ S_2 = \bigcup_{i=1}^{L} I_{2i},$$
$$m_r = \max_{i \in S_r} X_{i,h},\ r = 1, 2.$$

Then $N_h = \max(m_1, m_2)$.

Notice that, given $\{n_i\}$, m_1 and m_2 are independent. Consequently,

$$\mathbb{P}(N_h < k|\{n_i\}) = \mathbb{P}(m_1 < k|\{n_i\})\mathbb{P}(m_2 < k|\{n_i\}). \tag{14}$$

The idea is to find the conditional distributions of m_1 and m_2 given $\{n_i\}$ and then to average over the joint distribution of $\{n_i\}$.

We give here only the derivation of $\mathbb{P}(m_1 < k|\{n_i\})$, the conditional distribution of m_2 is found analogously.

For $i = 1, ..., L+1$ and $t = 1, ..., b$ we will denote

$$Y_i(t) := X_{(i-1)h+1,t}$$

the number of 1's in the first t trials of the i-th odd-numbered group I_{2i-1}. Note that $Y_i(b) = n_{2i-1}$.

The key observation is that, given $\{n_i\}$, $m_1 < k$ provided that $\{n_i\}$ is in the set Q_k defined in the statement, and further that

$$X_{(i-1)h+t+1,h} < k, \text{ for all } t = 1, ..., b-1, \ i = 1, ..., L.$$

But

$$X_{(i-1)h+t+1,h} = Y_{i+1}(t) + n_{2i} + n_{2i-1} - Y_i(t),$$

thus we can write

$$\mathbb{P}(m_1 < k|\{n_i\}) = \mathbb{P}(\bigcap_{t=1}^{b}\bigcap_{i=1}^{L}\{Y_i(t) + \alpha_i > Y_{i+1}(t) + \alpha_{i+1}\}|\{n_i\}), \tag{15}$$

where

$$\alpha_i - \alpha_{i+1} = k - n_{2i-1} - n_{2i} > 0, \ i = 1, ..., L$$

and hence

$$\alpha_i := (L - i + 1)k - \sum_{j=2i-1}^{2L} n_j, \ i = 1, ..., L+1. \tag{16}$$

The probability in the right-hand side of (14) appears in a variant of the L-candidate ballot problem (see Naus [16]).

Using the relation (2.5) from Naus [16] we obtain

$$\mathbb{P}(m_1 < k|\{n_i\}) = \det|h_{ij}|, \tag{17}$$

where, for $i, j = 1, ..., L+1$:

$$h_{ij} = \frac{n_{2i-1}!(b - n_{2i-1})!}{(n_{2i-1} + \alpha_i - \alpha_j)!(b - n_{2i-1} - \alpha_i + \alpha_j)!}, \tag{18}$$

with the convention $h_{ij} = 0$ if any of the factorial terms is negative.

From (15), (16) and (17) we deduce that, for $\{n_i\}$ in Q_k, we have

$$\mathbb{P}(m_1 < k|\{n_i\}) = R \det|d_{ij}^{(k)}|, \tag{19}$$

where

$$R = \prod_{i=1}^{L+1} n_{2i-1}!(b - n_{2i-1})!$$

and $d_{ij}^{(k)}$ are as given in the statement.

In a similar way we can show that

$$\mathbb{P}(m_2 < k | \{n_i\}) = T \det |g_{ij}^{(k)}|, \tag{20}$$

where

$$T = \prod_{i=1}^{L} n_{2i}!(h - b - n_{2i})!$$

and $g_{ij}^{(k)}$ are as in the statement.

By substituting (18) and (19) into (13) and then averaging over the distribution of the n_i's, given in (12), the formula in the statement follows. □

From relation (11), Proposition 6 and Proposition 7 we deduce the following result about the distribution of $d^*(\pi_{i,h}, Id_n)$.

Theorem 8. *For $0 \leq d \leq h-1$, we have*

$$\mathbb{P}(d^*(\pi_{i,h}, Id_n) \leq d) = \frac{1}{\binom{h}{\ell}} \sum_{\ell=h-d}^{h} \mathbb{P}(L^* = \ell) \sum_{m=h-d}^{\ell} \binom{h-m-1}{h-\ell-1} p(m, h-d),$$

where

$$\mathbb{P}(L^* = \ell) = \mathbb{P}(N_h < \ell + 1 | X_{1,n} = h) - \mathbb{P}(N_h < \ell | X_{1,n} = h)$$

and the two conditional probabilities are given by Proposition 7.

Note that the values for n and h which are typical for our application are not too large (n is of the order of 100) and hence there is no problem in computing the distribution of N_h given by Proposition 7.

6. Conclusion

Among the three "distances" presented in this article, we think that from the biological point of view the first one and the third one are the most interesting.

Our result on the distribution of the second distance may however have an interest in itself, mathematically speaking, or one may find some better applications to other problems.

While the third distance is specifically adapted to the reference region approach, the first distance could also be used in whole genome comparisons or window sampling approaches.

Based on the first idea, of comparing only the order of the orthologs which are in common between the two clusters, one could imagine replacing the transposition distance with another distance, maybe more relevant biologically. For

example, we could use the block-interchange distance of Christie [4] and the results of Doignon and Labarre [8] on its distribution.

A natural continuation of this work would be to try to obtain analogous results in the case of signed permutations (when we take into account also genes' orientation) or in the case of multipermutations (when we can have multiple orthologs in the genome B for a given gene in the reference region, as a result of duplication events occurred after the speciation of the two species). The case of multichromosomal genomes could also be considered.

Another important question to be considered is how to cleverly combine, in a single statistical test, the proximity of the orthologs and their order.

Acknowledgements

This work was partially supported by the ANR MAEV under contract ANR-06-BLAN-0113.

I would like to thank Etienne Pardoux, my thesis advisor, for many helpful discussions during this work, and Pierre Pontarotti, my second thesis advisor, for interesting biological discussions.

I also wish to thank Anthony Labarre for explaining to me his results on the number of cycles in the breakpoint graph.

I am grateful to the referee for his/her helpful suggestions, which have contributed to the improvement of the presentation of this article and have provided interesting ideas for future work.

References

[1] N. Berestycki, R. Durrett, A phase transition in the random transposition random walk, Probab. Theory Relat. Fields 136 (2006) 203 - 233.

[2] A. Bergeron, S. Corteel, M. Raffinot, The algorithmic of gene teams, Lecture Notes in Comput. Sci. 2452 (2002) 464 – 476.

[3] B. Bollobas, Random Graphs, Cambridge University Press, Cambridge, 2001.

[4] D.A. Christie, Sorting permutations by block-interchanges, Inform. Process. Lett. 60 (1996) 165 – 169.

[5] L. Comtet, Advanced Combinatorics: The Art of Finite and Infinite Expansions, Reidel Publishing Company, Dordrecht-Holland, 1974.

[6] E. Danchin, P. Pontarotti, Statistical evidence for a more than 800-million-year-old evolutionarily conserved genomic region in our genome, J. Mol. Evol. 59 (2004) 587 – 597.

[7] P. Diaconis, Group Representations in Probability and Statistics, Institute of Mathematical Statistics, Hayward, 1988.

[8] J-P. Doignon, A. Labarre, On Hultman numbers, J. Integer Seq. 10 (2007) Article 07.6.2.

[9] D. Durand, D. Sankoff, Tests for gene clustering, J. Comput. Biol. 10 (2003) 453 – 482.

[10] N. Eriksen, A. Hultman, Estimating the expected reversal distance after a fixed number of reversals, Adv. Appl. Math. 32 (2004) 439 – 453.

[11] J. Glaz, J. Naus, S. Wallenstein, Scan Statistics, Springer Verlag, New York, 2001.

[12] R. Hoberman, D. Durand, The incompatible desiderata of gene cluster properties, Lect. Notes in Bioinform. 3678 (2005) 73 – 87.

[13] R. Hoberman, D. Sankoff, D. Durand, The statistical analysis of spatially clustered genes under the maximum gap criterion, J. Comput. Biol. 12 (2005) 1083 – 1102.

[14] R.J. Huntington, J. Naus, A simpler expression for kth nearest neighbor coincidence probabilities, Ann. Probab. 3 (1975) 894 – 896.

[15] Z. Li, L. Wang, K. Zhang, Algorithmic approaches for genome rearrangement: a review, IEEE Transactions on Systems, Man and Cybernetics, Part C, 36 (2006) 636 – 648.

[16] J. Naus, Probabilities for a generalized birthday problem, J. Amer. Statist. Assoc. 69 (1974) 810 – 815.

[17] N. Raghupathy, D. Durand, Individual gene cluster statistics in noisy maps, Lect. Notes in Bioinform. 3678 (2005) 106 – 120.

[18] D. Sankoff, L. Haque, Power boosts for clusters tests, Lecture Notes in Comput. Sci. 3678 (2005) 121 – 130.

[19] D. Sankoff, L. Haque, The distribution of genomic distance between random genomes, J. Comput. Biol. 13 (2006) 1005 – 1012.

[20] W. Xu, D. Sankoff, Tests for gene clusters satisfying the generalized adjacency criterion, Lect. Notes in Bioinform. 5167 (2008) 152 – 160.

[21] W. Xu, C. Zheng, D. Sankoff, Paths and cycles in breakpoint graphs of random multichromosomal genomes, Lect. Notes in Bioinform. 4205 (2006) 51 – 62.